Reconnaissance de sauts d'athlète dans les vidéos : fusion du mouvement de caméra et de la trajectoire de points caractéristiques.

Emmanuel RAMASSO, Denis PELLERIN, Michèle ROMBAUT

Laboratoire des Images et des Signaux
LIS - INPG, 46 Avenue Félix Viallet, 38031 Grenoble, France
{Prénom.Nom}@lis.inpg.fr

Résumé – La reconnaissance de l'activité humaine est un champ de recherches très actif dans la communauté du traitement d'image. Dans cet article, nous nous intéressons à la reconnaissance de l'action vue en tant que composante de l'activité. Pour cela, nous utilisons une méthode de fusion de données basée sur la théorie des fonctions de croyances et plus précisément le Modèle de Croyance Transférable. Les données utilisées sont, d'une part, l'estimation du mouvement de la caméra et, d'autre part, la position de points caractéristiques de la silhouette humaine. La méthode est testée pour reconnaître les actions *course*, *saut* et *chute* d'un athlète dans des vidéos de meeting d'athlétisme contenant diverses séquences de saut.

Abstract – Human activity recognition is an active field of research in image understanding and computer vision. We pay attention to action recognition, as an atomic part of an activity. The method is based on data fusion using belief theory and more precisely the Transferable Belief Model. Data concern the camera motion and the trajectory of feature points. The developed method is tested on videos of athletics meetings in order to recognize *running*, *jumping* and *falling* actions in different video sequences of jump.

1 Introduction

L'analyse du mouvement humain dans les vidéos [1] est un challenge important dans le domaine de la vision par ordinateur. Les applications sont nombreuses notamment pour la surveillance de lieux publics ou privés, la recherche et l'archivage de vidéos, le diagnostic médical ou encore l'analyse sportive.

La finalité d'une analyse de mouvement humain est de reconnaître une *activité* que nous percevons comme une *séquence d'actions*. Dans cet article, nous nous focalisons sur la reconnaissance d'actions en vue de celle de l'activité.

« Reconnaître » nécessite de comparer une observation à des modèles et de choisir le plus approprié. De nombreuses méthodes proposées dans la littérature sont basées sur les probabilités [2] avec particulièrement les chaînes de Markov cachées (HMM) et les réseaux bayésiens dynamiques (DBN), ces deux méthodes étant décrites et utilisées dans [3].

La reconnaissance d'actions nécessite d'extraire, à partir des vidéos, des attributs pertinents. Ces attributs sont généralement dépendants de l'application et proviennent dans la majorité des cas d'algorithmes de suivi de points caractéristiques [4] de la silhouette. Quelle que soit leur provenance, les attributs sont généralement entachés d'*imprécisions* dues aux capteurs (e.g. la caméra) ou aux algorithmes de traitement des données. Classiquement, la théorie des probabilités est utilisée pour prendre en compte ces *imperfections*. Nous avons choisi le Modèle de Croyance Transférable (TBM) proposé par Smets et Kennes [5] car ce modèle, en plus de prendre en compte l'imprécision des données, explicite le *doute* ainsi que le *conflit* entre des sources d'informations. Le TBM semble bien adapté à la reconnaissance des actions : d'une part, le doute permet d'exprimer les situations d'*incertitude* entre des hypothèses ce qui s'avère fort intéressant pour décrire l'*aspect graduel des transitions* entre les actions et, d'autre part, le conflit *quantifie* la nécessité d'améliorer la modélisation des informations fusionnées et peut aussi remettre en cause la *qualité* des données.

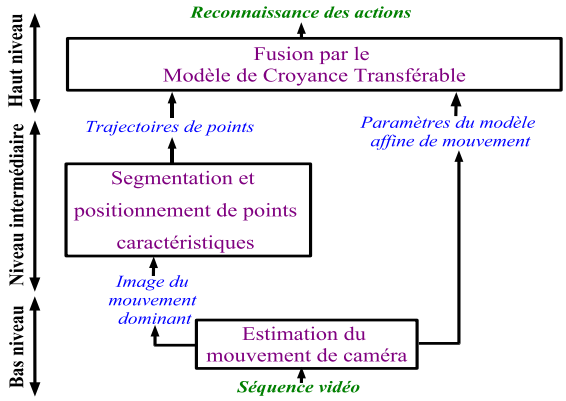

FIG. 1: Architecture du système de reconnaissance d'actions.

Une méthode originale basée sur le TBM est proposée pour le reconnaissance des actions d'athlètes dans des vidéos acquises avec caméra mobile. Une architecture en trois niveaux est présentée dans la section 2. La section 3 est consacrée à la description des deux premiers niveaux qui fournissent les paramètres utilisés par le module haut niveau. Ce dernier, décrit dans la section 4, réalise la fusion et la reconnaissance. Le système est testé sur des vidéos d'athlétisme où l'objectif est de retrouver les actions *course*, *saut* et *chute* dans quatre types de sauts différents. Les résultats présentés dans la section 5 montrent la validité de la méthode.

2 Architecture du système

Le synoptique du système est représenté figure 1. Le fonctionnement est basé sur deux hypothèses liées à l'application :

TAB. 1: Paramètres des niveaux bas et intermédiaire.

Mouvement de caméra		Les points	
a_0	horizontal	(x_c, y_c)	centre de gravité
a_1	vertical	(x_h, y_h)	tête
a_2, a_5	divergence	(x_l, y_l)	pieds

(i) la caméra suit le mouvement de l'athlète, (ii) les mouvements dans l'image de la tête, du centre de gravité et d'un pied de l'athlète renseignent sur l'action en cours. A partir de ces deux hypothèses, nous proposons une structure à trois niveaux, chacun d'eux correspondant à un niveau d'abstraction des données traitées. Tout d'abord, partant de la vidéo originale (fig. 2(a)), un module *bas niveau* estime le mouvement de caméra et fournit, d'une part, six paramètres correspondant au modèle affine et, d'autre part, une image dont l'intensité d'un pixel reflète son appartenance au mouvement dominant. Ensuite, au *niveau intermédiaire*, le positionnement des points caractéristiques sur la silhouette préalablement segmentée est réalisé. Enfin, au *haut niveau*, le mouvement de caméra et la position des points sont combinés dans le cadre du TBM pour obtenir la croyance sur la réalité des actions.

3 Paramètres caractérisant les actions

3.1 Le mouvement de caméra

Nous supposons que la caméra suit grossièrement l'athlète dont les mouvements sont analysés. Nous avons utilisé la méthode robuste d'estimation de modèles paramétriques de mouvement présentée dans [6]. Cette méthode itérative et multi-résolution utilise deux images successives. Le modèle affine (eq. 1) est généralement suffisant pour un grand nombre d'application :

$$\begin{cases} v_x = a_0 + a_2 \cdot x + a_3 \cdot y \\ v_y = a_1 + a_4 \cdot x + a_5 \cdot y \end{cases} \quad (1)$$

Les paramètres a_i utilisés par le module de reconnaissance sont regroupés dans la table 1. Ces paramètres correspondent au mouvement de caméra avec : les translations horizontale et verticale (a_0 et a_1), le zoom (a_5 et a_2) et la rotation (a_4 et a_3 non utilisés). Un lissage de ces paramètres est réalisé par un filtrage Gaussien. Ces paramètres permettent de connaître l'appartenance de chaque pixel au mouvement dominant qui correspond généralement au mouvement du fond (fig. 2(b)).

3.2 Positionnement de points caractéristiques

La méthode décrite dans [7] est exploitée afin de positionner trois points caractéristiques sur une silhouette : deux sont situés aux extrémités (assimilées à la tête et à l'extrémité d'une jambe), et le troisième correspond au centre de gravité (tronc). La trajectoire de ces points est supposée suffisamment pertinente pour la reconnaissance d'actions globales telles qu'une *course*, un *saut* ou une *chute*. Préalablement au positionnement, il est nécessaire de séparer la silhouette du fond.

Segmentation de la silhouette : La segmentation est basée sur l'image du mouvement dominant, obtenue par l'analyse du mouvement de caméra. Un filtrage médian et des opérations morphologiques permettent d'isoler la silhouette (fig. 2(c)).

TAB. 2: Méta-paramètres.

paramètres bruts	méta-paramètres
a_0, a_2	mouvement horizontal
a_1	mouvement vertical
$(x_c, y_c), (x_h, y_h), (x_l, y_l)$	« swing »
$(x_c, y_c), (x_h, y_h), (x_l, y_l)$	alternance des jambes
y_c	variation verticale

Positionnement des points caractéristiques : Le centre de gravité de la silhouette segmentée est calculé et ce premier point caractéristique est ensuite utilisé pour déterminer l'orientation de l'axe principal de la silhouette. Les deux points extrêmes de cette dernière sont alors déduits et correspondent aux deux autres points caractéristiques (fig. 2(c)). Une procédure de re-classification des pixels de la silhouette est réalisée à chaque image afin de rendre plus robuste le positionnement : le principe est de seuiller la plus petite des distances entre les pixels de la silhouette et les points caractéristiques obtenus à l'image précédente puis d'affecter les pixels trop éloignés au fond.

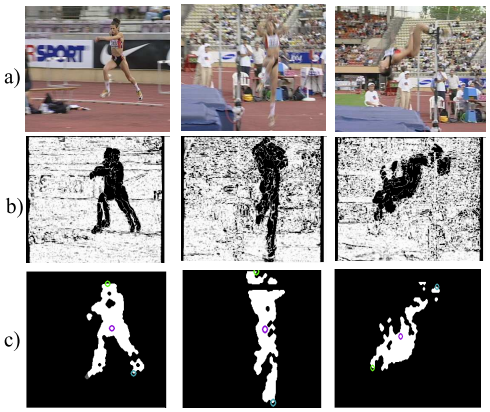

FIG. 2: Exemple de résultat pour un saut en hauteur : (a) images originales pour les actions *course*, *saut* et *chute*, (b) images du mouvement dominant et (c) segmentation et positionnement des points caractéristiques.

3.3 Méta-paramètres

Afin de faciliter la description des actions, des paramètres plus évolués (tab. 2) sont calculés à partir de ceux obtenus par les algorithmes (tab. 1). Par exemple, l'angle que fait l'axe de la silhouette de l'athlète avec l'horizon (« swing ») est calculé à partir des coordonnées des points caractéristiques et permet d'avoir une information sur la manière dont est positionné l'athlète dans l'espace (fig. 3).

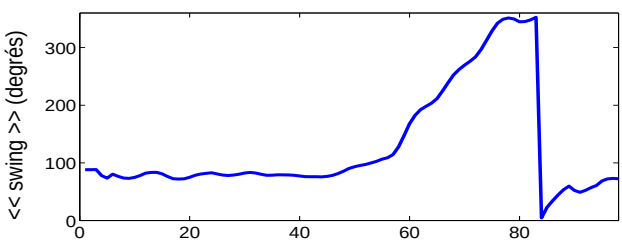

FIG. 3: Angle entre l'axe de la silhouette et l'horizon (« swing ») pour un saut en hauteur.

4 Fusion et reconnaissance

L'évolution dans le temps des différents paramètres obtenus aux deux précédents niveaux doit permettre de reconnaître l'action en cours. Nous proposons une structure de fusion de données basée sur le Modèle de Croyance Transférable (TBM). A chaque valeur des paramètres est associée une croyance en la réalité des actions possibles. Ces croyances sont fusionnées dans le cadre du TBM afin d'obtenir une croyance plus complète tenant compte de l'ensemble des paramètres.

4.1 Le Modèle de Croyance Transférable

Le TBM a été proposé par Smets et Kennes dans [5] et fait suite aux travaux de Dempster et Shafer sur la théorie de l'évidence. Le TBM offre des outils pour la combinaison de sources d'évidence, chacune d'elles exprimant une croyance sur des hypothèses concernant le monde dans lequel elle se trouve.

L'espace de discernement regroupe les différentes hypothèses possibles. Dans le cadre de cet article, l'espace de discernement est $\Omega_A = \{V_A, F_A\}$ et représente l'ensemble des hypothèses concernant une action A, où V_A correspond à « *A est vraie* » et F_A signifie « *A est fausse* ». Les hypothèses sont exclusives et l'espace de discernement est exhaustif.

La masse d'évidence $m^{\Omega_A}_{S_i}$ représente la quantité de croyance affectée à chaque partie (formée par les hypothèses et unions d'hypothèses) de Ω_A par une source S_i. Pour chaque image d'une vidéo, une telle affectation est réalisée à partir de la valeur de chaque paramètre. Cette distribution de masses est réalisée par une application définie $\forall X \in 2^{\Omega_A}$ à valeur dans [0,1] où $2^{\Omega_A} = \{V_A, F_A, V_A \cup F_A\}$. Par construction, $m^{\Omega_A}_{S_i}(\emptyset) = 0$. L'ensemble $V_A \cup F_A$ représente le doute sur la véracité de l'action A sans privilégier une des deux hypothèses.

L'allocation de la masse d'évidence est réalisée pour chaque paramètre mesuré, et relativement aux actions, par une méthode inspirée des sous ensembles flous et revient à une conversion numérique-symbolique. Une illustration est donnée à la figure 4, où a_0 représente le mouvement de translation horizontale de la caméra et où les paramètres de règles ont été définis pour une action *course*. Si $a_0 = 7$ (pixels/image) alors on a $m^{\Omega_A}_{a_0}(V_A) = 1$ (croyance catégorique) et cela signifie que la proposition « A = *course* » est vraie. Si $a_0 = 2,3$ alors $m^{\Omega_A}_{a_0}(F_A) = 0,67$ et $m^{\Omega_A}_{a_0}(V_A \cup F_A) = 0,33$ ce qui signifie que l'action *course* est plutôt fausse mais qu'il persiste un doute. La difficulté est de régler les paramètres de conversion qui sont pour le moment fixés par heuristiques. Cette difficulté est récurrente quel que soit le formalisme adopté (probabilité, évidence, possibilité) si l'on ne dispose pas d'une base d'apprentissage suffisamment riche comme c'est le cas ici.

La combinaison des distributions de masses permet de déterminer la véracité des actions prenant en compte l'ensemble des paramètres. Plusieurs règles de combinaisons ont été proposées, par exemple, la combinaison conjonctive, notée $\textcircled{\cap}$, définie par :

$$(m^{\Omega_A}_{S_1} \textcircled{\cap} m^{\Omega_A}_{S_2})(X) = \sum_{\substack{X,Y,Z \subseteq \Omega_A \\ Y \cap Z = X}} m^{\Omega_A}_{S_1}(Y).m^{\Omega_A}_{S_2}(Z) \qquad (2)$$

avec S_1 et S_2 deux sources distinctes et toutes deux définies sur le même espace de discernement Ω_A. La règle $\textcircled{\cap}$ étant commutative et associative, il est possible de combiner n sources distinctes S_i en appliquant cette règle en cascade et on obtient $m^{\Omega_A}_{S_{1,2\ldots n}} = m^{\Omega_A}_{S_1} \textcircled{\cap} m^{\Omega_A}_{S_2} \ldots \textcircled{\cap} m^{\Omega_A}_{S_n}$. La règle $\textcircled{\cap}$ est utilisée lorsque plusieurs sources doivent concorder concernant la véracité de A. Lorsque des sources sont en conflit, de la masse apparaît sur l'ensemble vide ($m^{\Omega_A}_{S_{1,2\ldots n}}(\emptyset) > 0$) et ce conflit doit être analysé et résolu car il est absorbant par la règle $\textcircled{\cap}$.

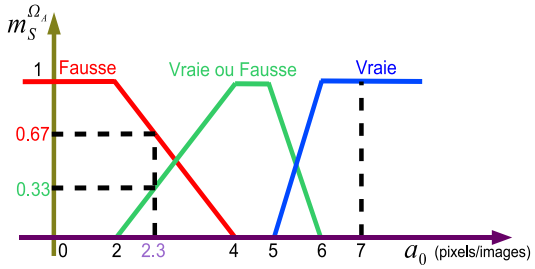

FIG. 4: Un exemple d'allocation de masse pour le paramètre de mouvement de translation horizontale a_0.

4.2 Reconnaissance d'action

La reconnaissance d'une action A est réalisée en trois étapes. Tout d'abord, les distributions de masses, définies sur 2^{Ω_A}, sont déterminées (fig. 4) pour chaque paramètre, les seuils des conversions numériques-symboliques étant fixés pour chaque action et pour chaque type de saut. Ensuite, la combinaison des distributions est réalisée (eq. 2). Enfin, des contraintes temporelles sont appliquées sur la durée minimale des actions : si une action A devient vraie après combinaison, c'est à dire que $m^{\Omega_A}_{S_{1,2\ldots n}}(V_A) > 0$, alors un changement potentiel dans le comportement est détecté. Si cette durée est inférieure à un seuil Δ_A alors la masse est transférée sur le doute, $V_A \cup F_A$, assimilant $m^{\Omega_A}_{S_{1,2\ldots n}}(V_A)$ à du bruit. Ce processus agit comme un filtre passe bas.

Les actions sont indépendantes c'est à dire que la croyance sur la véracité d'une action n'influe pas sur la croyance concernant la véracité des autres actions. De plus, les actions peuvent être vraies en même temps puisqu'elles ont été définies les unes indépendamment des autres (non exclusives). Ce choix a été fait pour pouvoir poursuivre le travail présenté dans cet article vers la reconnaissance d'activités. En effet, une activité étant vue comme une séquence d'actions, une transition entre deux actions successives est une zone (ensemble d'images) où les deux actions peuvent être vraies en même temps.

5 Expérimentations

Conditions expérimentales : La méthode est testée pour la reconnaissance des actions *course*, *saut* et *chute* dans quatre types de saut différents et sur 33 vidéos (tab. 3). Les vidéos, au format 290 × 292 pixels, présentent des angles de vue différents et *a priori* inconnus. Généralement, un seul athlète bouge dans la vidéo mais il arrive que d'autres personnes soient en mouvement en même temps puisque ces vidéos sont des extraits de meetings d'athlétisme. Le réglage des paramètres des conversions numériques-symboliques est réalisé une fois pour chaque action dans chaque type de saut. Enfin, le système de reconnaissance fonctionne, pour le moment, image par image.

Décision et évaluation : Les résultats en sortie du module de reconnaissance sont sous la forme de masses d'évidence.

TAB. 3: La base de vidéos : le nombre d'images concernant les actions *course*, *saut* et *chute* (col. 3-5) et le nombre total de vidéos pour chaque saut (N_V).

saut/action	N_V	course	saut	chute	total
hauteur	9	604	351	205	1160
longueur	8	632	220	213	1065
perche	8	598	417	243	1258
triple saut	8	576	505	377	1458
total	33	2410	1493	1038	4941

TAB. 4: Rappels et précisions (en %) de la reconnaissance des actions *course*, *saut* et *chute* basée sur leur crédibilité.

saut/action	course		saut		chute	
	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$
hauteur	$96,4$	$79,6$	$75,4$	$73,3$	$74,1$	$91,6$
longueur	$90,0$	$80,6$	$58,2$	$52,9$	$64,8$	$70,8$
perche	$81,9$	$75,6$	$74,5$	$70,9$	$72,8$	$75,6$
triple saut	$85,9$	$70,6$	$55,6$	$63,3$	$62,9$	$52,1$
total	$88,6$	$76,7$	$59,6$	$66,1$	$67,8$	$64,0$

Aucune prise de décision n'est volontairement réalisée car ces données alimenteront un module de reconnaissance d'activités. Cependant, l'évaluation de la reconnaissance d'actions nécessite de savoir si une action est vraie ou fausse. Pour cela, nous considérons que si la crédibilité d'une action est non nulle ($m^{\Omega_A}_{S_{1,2\ldots n}}(V_A) > 0$) alors l'action est vraie (fig. 5). C'est une décision stricte par rapport à celle qui considérerait la plausibilité ou la probabilité pignistique. On utilise les indices de rappel et de précision ($\mathcal{R}$ et $\mathcal{P}$) pour évaluer les résultats avec $\mathcal{R}= \frac{C \cap R}{C}$ et $\mathcal{P}= \frac{C \cap R}{R}$ où C est l'ensemble de référence, R est l'ensemble des images retrouvées par le module de reconnaissance (relativement au critère basé sur la crédibilité) et $C \cap R$ est le nombre d'images correctement retrouvées.

Description et analyse des résultats : La table 4 regroupe les valeurs de $\mathcal{R}$ et $\mathcal{P}$. La dernière ligne est la moyenne des résultats sur toutes les vidéos. L'action *course* est convenablement reconnue avec $\mathcal{R} \geq 81.9\%$ et $\mathcal{P} \geq 70,6\%$. Quant aux actions *saut* et *chute*, on peut distinguer deux ensembles de résultats selon le type de saut : un premier pour le saut en hauteur et à la perche avec $\mathcal{R} \geq 72.8\%$ et $\mathcal{P} \geq 70.9\%$, et un second pour le saut en longueur et le triple saut avec $\mathcal{R} \geq 55.6\%$ et $\mathcal{P} \geq 52,1\%$. Les résultats sont dégradés par d'autres personnes ou objets en mouvement (e.g. la perche perturbe le positionnement des points). De plus, les actions ont été reconnues de manière statique mais une description tenant compte des dépendances temporelles entre les actions serait plus pertinente.

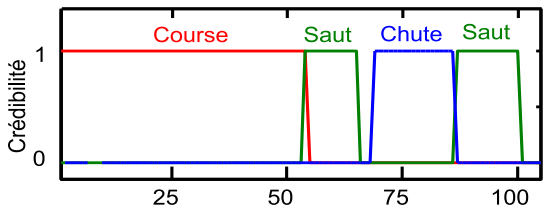

FIG. 5: Evolution de la crédibilité d'actions dans une séquence de saut en hauteur, le deuxième saut est dû à l'athlète qui exprime sa joie lorsqu'il réussit le saut.

6 Conclusion

Une méthode originale de reconnaissance de sauts d'athlète basée sur le Modèle de Croyance Transférable (TBM) a été proposée. Les données combinées proviennent d'un estimateur de mouvement de caméra ainsi que d'un module de positionnement de points caractéristiques de la silhouette. Les valeurs des paramètres ont été traduites en masses d'évidence obtenues grâce à un procédé inspiré des règles floues et fusionnées dans le cadre du TBM. Les tests ont porté sur 33 vidéos où l'objectif était de reconnaître les actions *course*, *saut* et *chute* dans quatre types de saut différents (longueur, hauteur, perche et triple saut). Les résultats ont permis de mettre en évidence l'intérêt de la méthode proposée qui se distingue de celles majoritairement employées dans la littérature, d'ailleurs généralement basées sur les probabilités. Un point fort de la méthode est la mise en évidence du conflit qui rend immédiatement compte d'un réglage inadapté des règles d'allocations de masses et permet de rectifier en conséquence. De plus, le doute entre les hypothèses permettra d'expliciter les transitions entre actions lors de la reconnaissance des activités. Enfin, la prise en compte d'informations complémentaires dans le processus de reconnaissance, comme des règles expertes ou d'autres *a priori*, est rapidement réalisable dans le cadre du TBM. Un point critique de la méthode concerne le réglage des paramètres permettant les allocations des masses : ne disposant pas d'une base d'apprentissage suffisante, nous visons à les adapter en ligne.

Remerciements

Cette recherche est en partie soutenue par le réseau d'excellence SIMILAR. Les auteurs remercient l'équipe Vista de l'Irisa/Inria Rennes pour le logiciel Motion2D sur lequel s'appuient les développements de cette étude.

Références

[1] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *PR*, 36(3):585–601, 2003.

[2] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition and probabilistic recognition methods. *CVIU*, 96:129–162, 2004.

[3] T. Xiang and S. Gong. Discovering bayesian causality among visual events in a complex outdoor scene. In *Proc. of the IEEE on Advanced Video and Signal based Surveillance*, pages 177–182, 2003.

[4] J. Wang and S. Singh. Video analysis of human dynamics-a survey. *Real-Time Imaging*, 9(5):321–346, 2003.

[5] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–234, 1994.

[6] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Vis. Comm. and Image R.*, 6(4):348–365, 1995.

[7] C. Panagiotakis and G. Tziritas. Recognition and tracking of the members of a moving human body. In *Articulated motion and deformable objects*, pages 86–98, 2004.